

# TPC DAM DAQ updates

Jin Huang (BNL)

Takao Sakaguchi (BNL)

Many thanks to discussion with Joe, Wei, Kai, Huchen, Martin, John and Ed

# Envelop parameters

**Proposed KPP:** demonstrate readout simulated data @ 1.43 Gbps x 400 fibers @ >99% LT

Input data stream:

400 GBT fibers total

Max continuous: 2.87 Gbps / fiber

Average continuous: 1.43 Gbps x 400 fibers

Clock/Trigger input:

Fiber, protocol TBD

Clock = 9.4 MHz

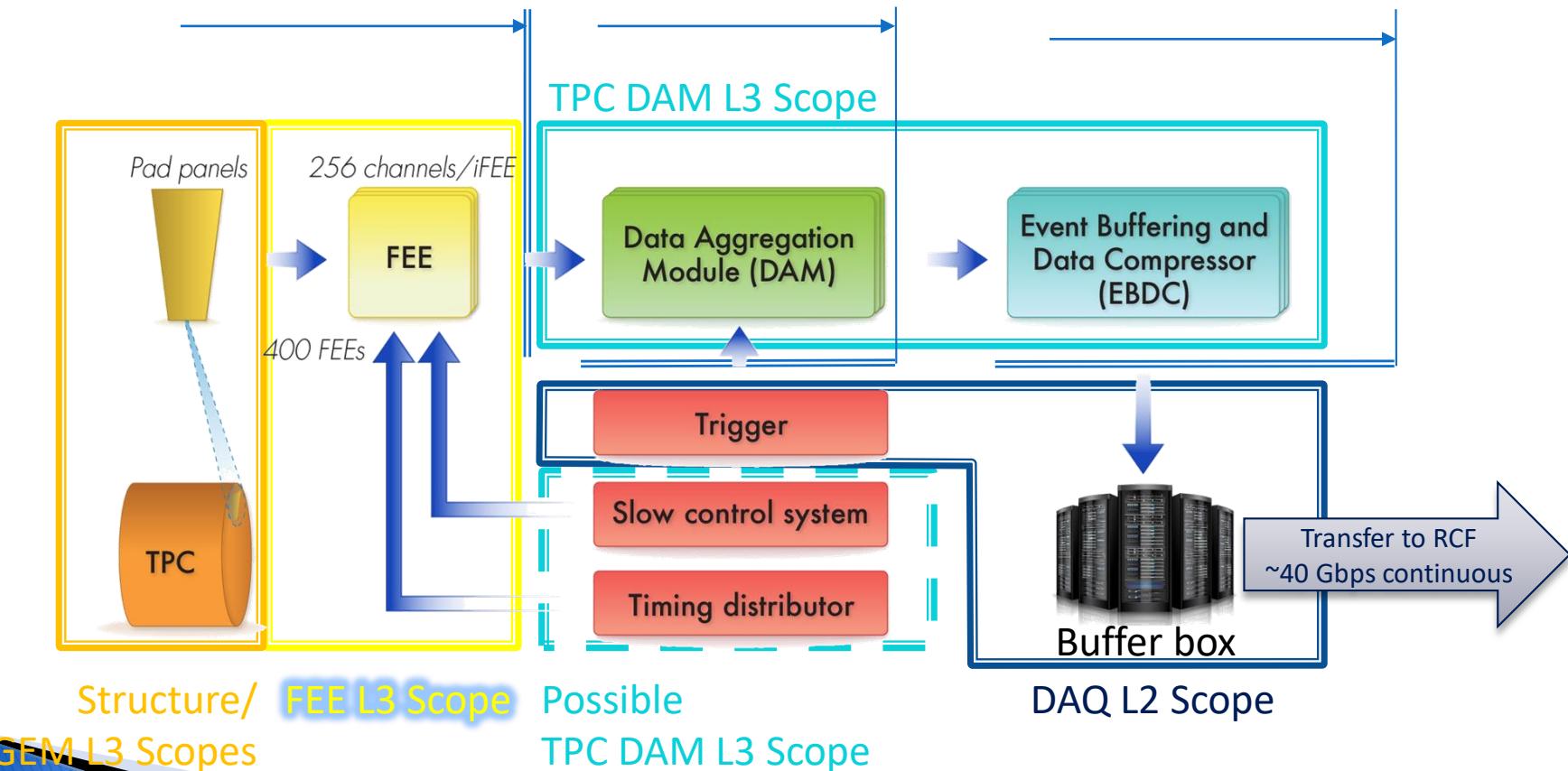
Trigger Rate = 15 kHz

Output data stream to buffer box:

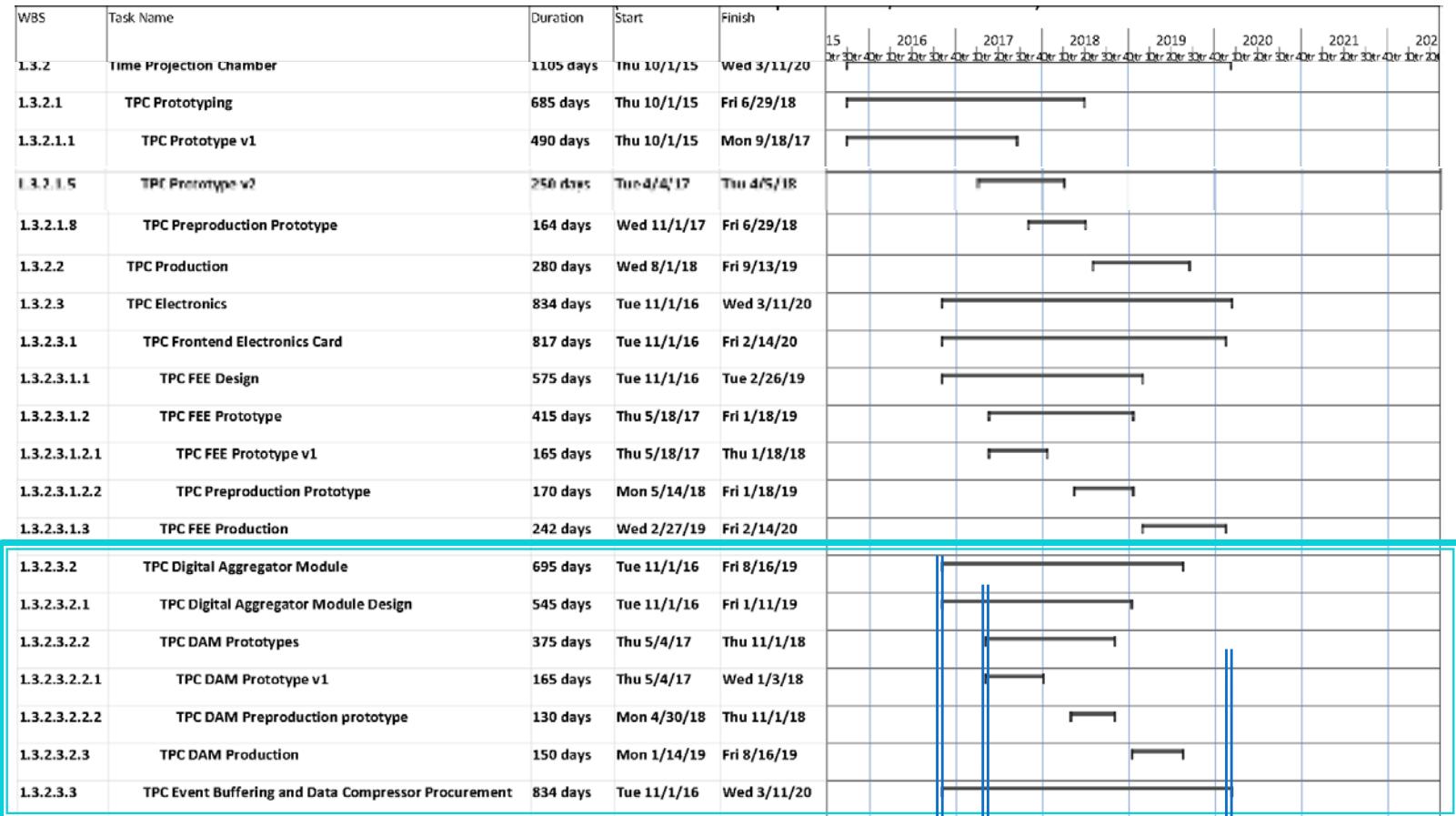
N x 10 Gbps Ethernet via fiber (N=10-50)

Total continuous limit: <120 Gbps (?)

i.e. 3x (Transfer rate to RCF ~ 40 Gbps)



# Timeline envelop. Cost <~ 0.45 M\$



Next Milestone

- Q4 2016, Design starts
- Apr 2016, Feasible design, BNL CD1 review
- Mid 2017, Prototype, 2 iterations possible
- Mid 2018, CD3-b authorization, production start
- Early 2020, Deliver all parts to 1008, establishing KPP
- Jan 2022, First beam

# ALICE TPC DAQ

JINST 11 (2016) C03021, ALICE TDR

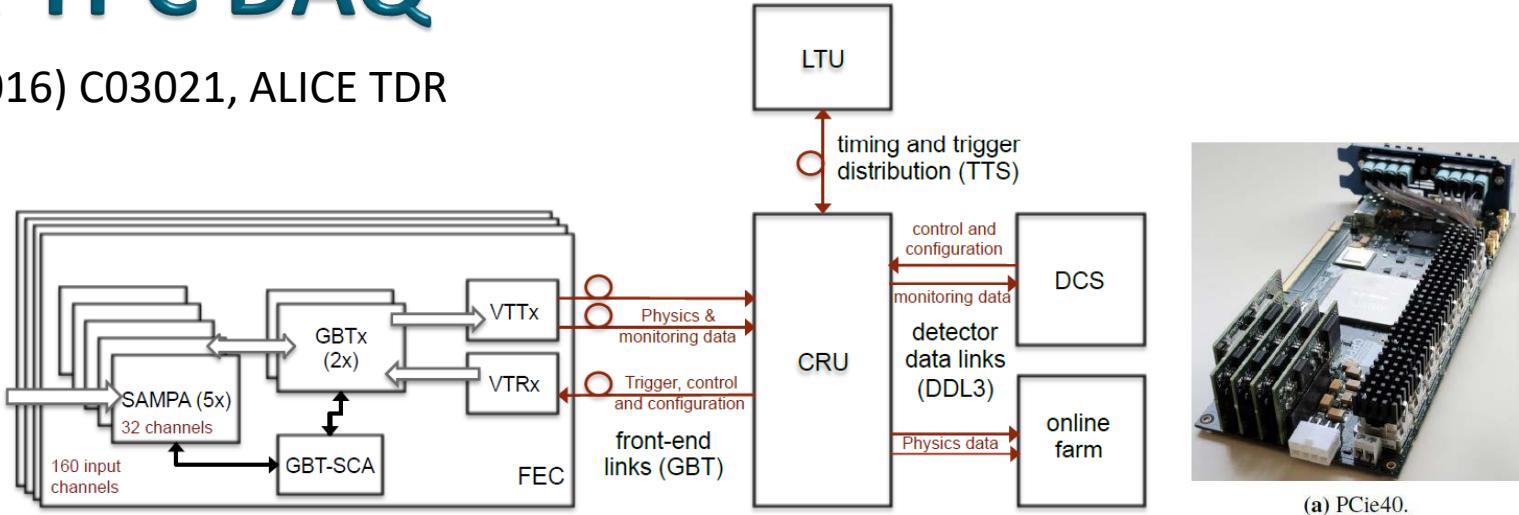
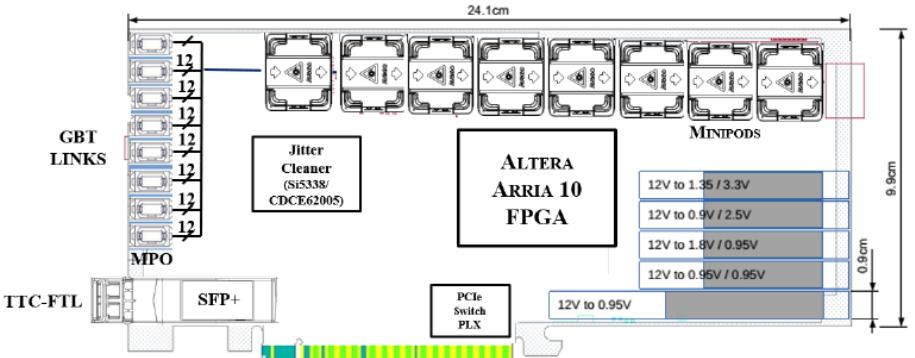


Figure 6.9: Schematic of the TPC readout system with the CRU as central part interfacing the front-end electronics to the trigger system, the DCS and the online farm.

## ALICE CRU based on LHCb PCIe40 card

- Prototyped by CPPM, Marseille, France
- Arria 10 family FPGA, (15K\$/chip?)
- 24 GBT input fibers [JINST 11 2016]
- PCIe Gen3 x16 interface
- TTC-FTL accepting timing/trigger
- Cost 15-20 k\$ (need to be confirmed)

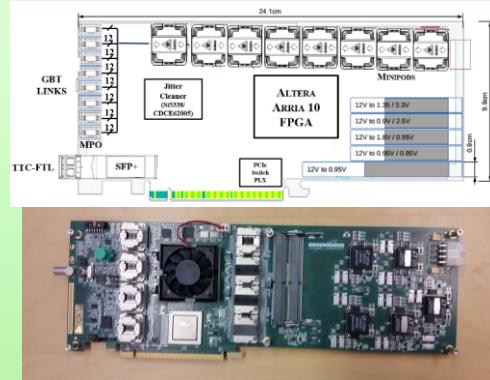


(b) PCIe40 Schematic.

# Our options: 10-50x (PCIe card + server)

Data Aggregation Module (DAM):  
PCIex8 or x16 card with multiple (8-48x) GBT fiber IO

Option 1: LHCb/ALICE CRU

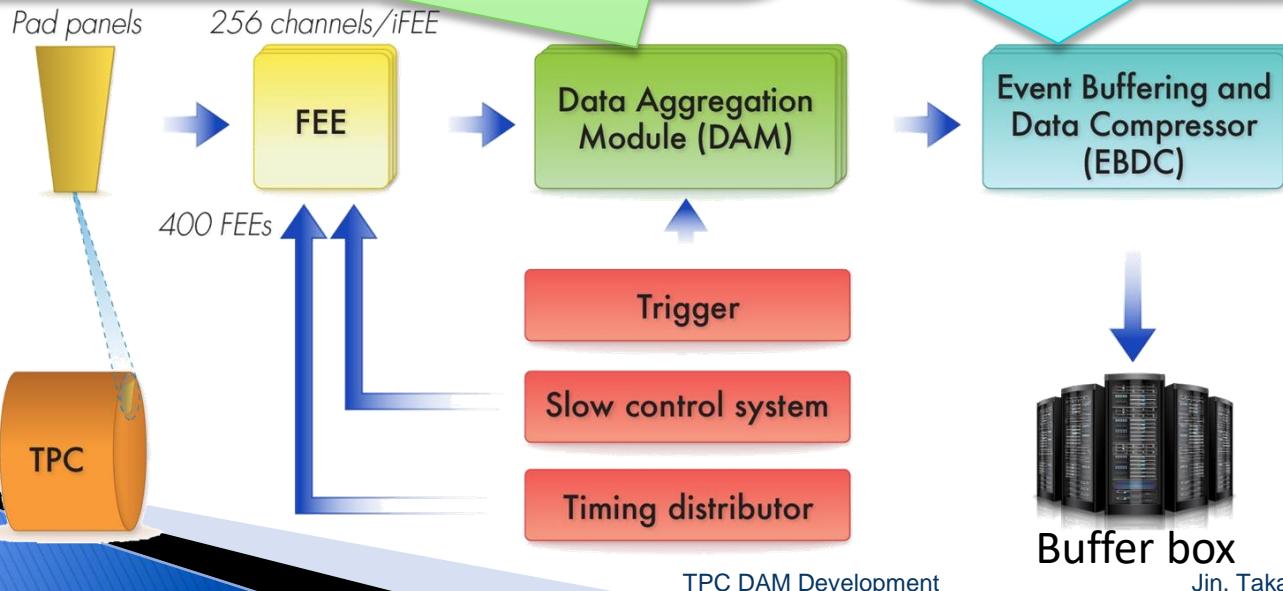


Option 2: ATLAS FELIX  
(see next talk)

Option 3: build our own based on ALICE/ATLAS exp.

Event Buffering and Data Compressor (EBDC): Rack server that can host at 1x PCIex16 cards + 2x 10 Gbps Ethernet port

Example: Dell PowerEdge R830  
2x12 cores, 2x10 GBps, ~ 10k\$



# FPGA Choices



(a) PCIe40.



FELIX



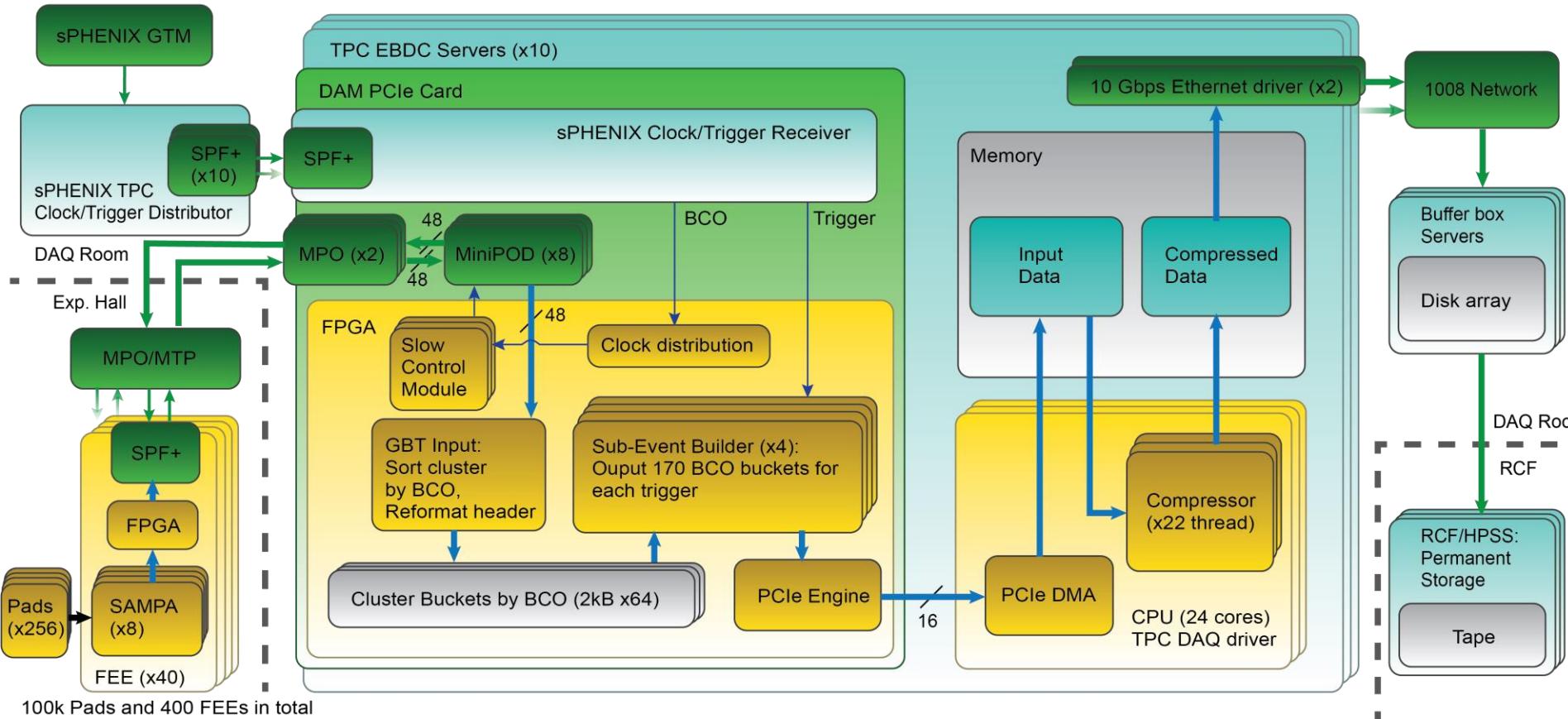
FPGA Family Name	Xilinx Virtex 6	Altera Stratix V GX	Xilinx Virtex 7	Altera Arria 10 GX **	Xilinx Virtex Ultrascale	Altera Stratix 10	CRU Requirements #	Xlinux Virtex Ultrascale
Status		available	available	ES available from Q2'15	available	end of 2017		Available
FPGA part number	XC6VLX240T	5SGXEA7	XC7VX690T	10AX115	XCVU190	10SG280		XCKU115
Used in	C-RORC	AMC40	MP7	PCIe40				FELIX v1.5 test boards
Logic Elements / Cells [M]	0.241	0.622	0.693	1.15	1.9	2.8		1.451
FFs [M]	0.3	0.939	0.866	1.7	2.14			1.3
LUTs [M]	0.15	0.235	0.433	0.425	1.07			0.66
18/20 Kb RAM Blocks	832	2560	2940	2713	7560	11721	1920 / 2560	4320
Total Block RAM (Mb)	15	50	53	53	133	229	40 / 53	75.9
≥ 10 Gb/s Transeivers	24	48	80	96	60	144	48	(48 input + 48 output fiber links in FELIX)
PLLs	12	28	20	32	60	48		48
PCIe x8, Gen3	2 (Gen2)	4	3	4	6	6		6

# TPC Detector is the majority user (>70%) of CRU boards. CRU requirements is measured against TPC detector specific logic occupancy.

\*\* Altough the maximum number of links of the Arria10 family is 96 links, the FPGA equiping the PCIe40 board has only 72 links

# Plausibility diagram

Assuming 10x (DAM + EBDC). We could afford more if bumping into bottle neck or be safe



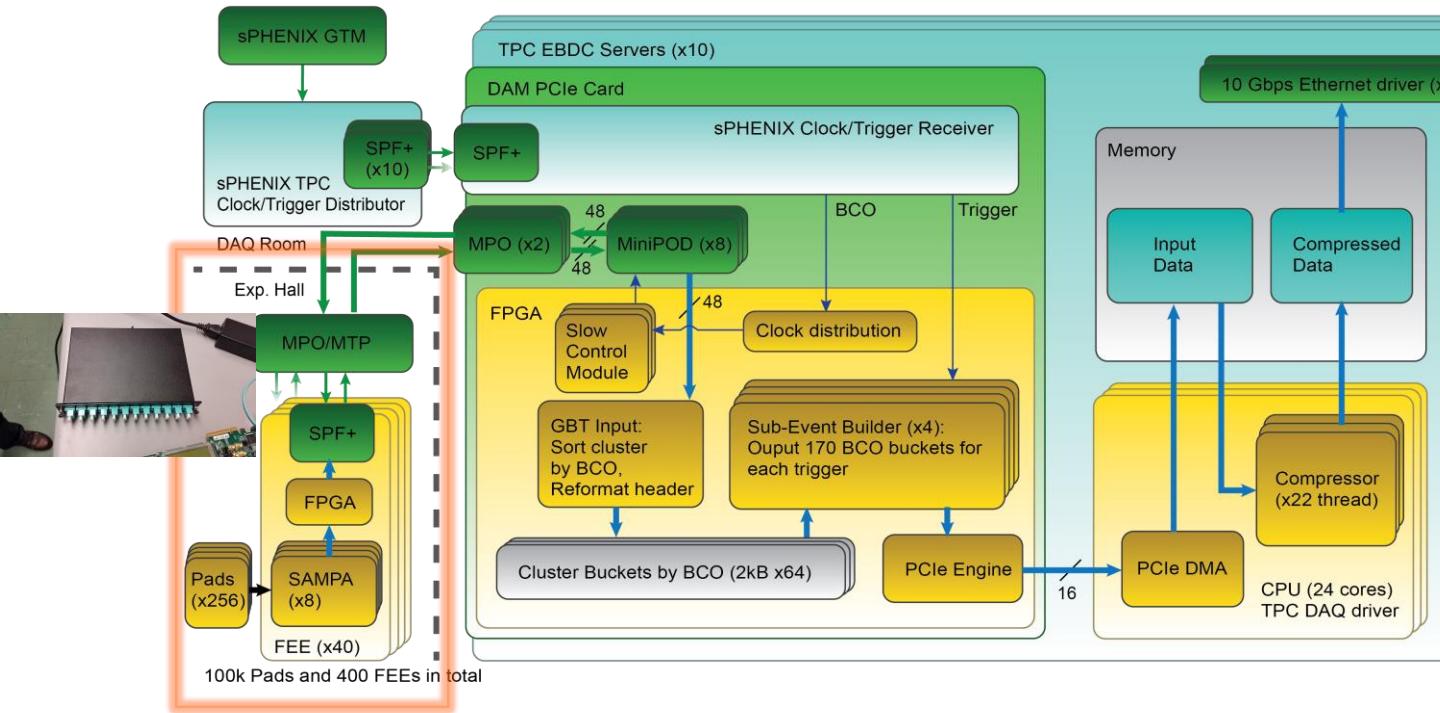
## Rate estimation spread sheets:

[https://docs.google.com/spreadsheets/d/1Q\\_uYf00\\_8pushSiYns29T\\_-ThIOqQaqpKbVS\\_LDqlAg/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1Q_uYf00_8pushSiYns29T_-ThIOqQaqpKbVS_LDqlAg/edit?usp=sharing)

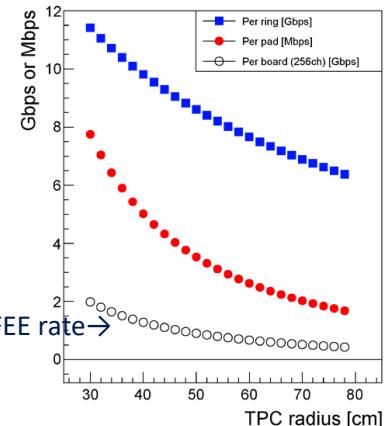
# Input stage

## Rate estimation spread sheets:

[https://docs.google.com/spreadsheets/d/1Q\\_uYf00\\_8pushSiYns29T\\_-ThlQqQaqpKbVS\\_LDqlAg/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1Q_uYf00_8pushSiYns29T_-ThlQqQaqpKbVS_LDqlAg/edit?usp=sharing)



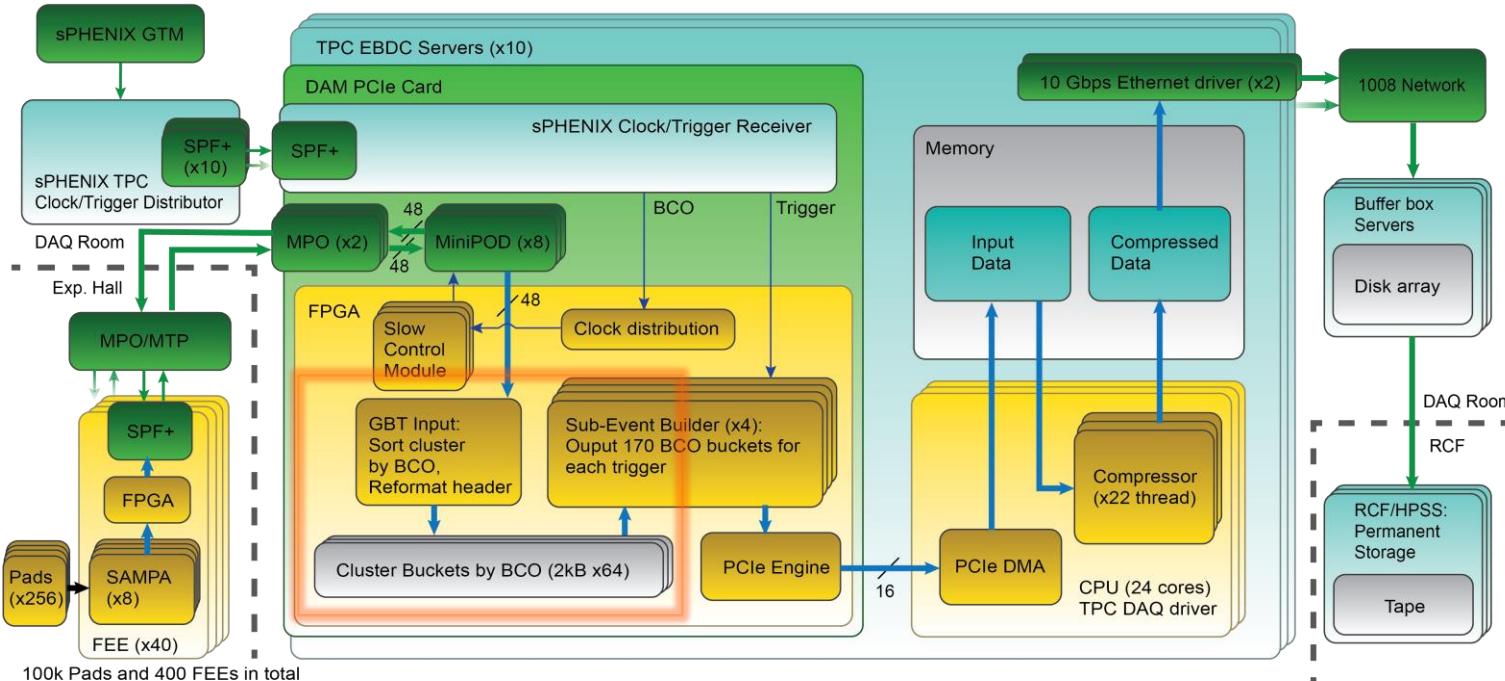
- ▶ Per DAM: 40 FEEs, each send data in 1 fiber
  - Data format in minimal chunk = one cluster in one channel: 2x10 bit header (channel ID + timing + length) + 5x10 bit wavelet
  - Wavelet sampled timed to BCO (beam collision clock = 9.4 MHz)
  - **Payload speed limit = 3.2 Gbps/fiber, 128 Gbps/DAM**
  - **Max continuous rate = 2.87 Gbps , 115 Gbps/DAM**
  - Average continuous rate = 1.43 Gbps
- ▶ Media: MTP fibers -> bundle to MPO. GBT/UPT protocol?
- ▶ Downlink fiber send clock and slow control to FEEs



# BCO buckets

## Rate estimation spread sheets:

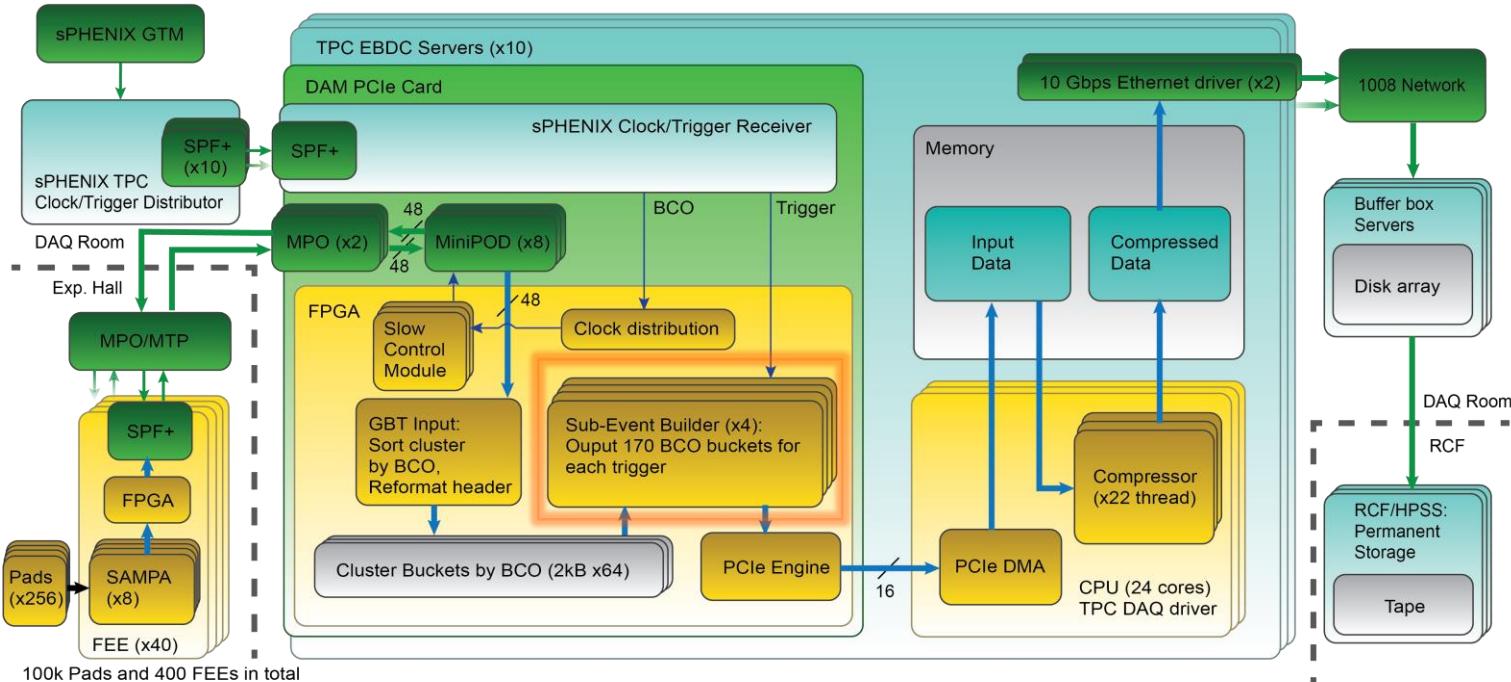
[https://docs.google.com/spreadsheets/d/1Q\\_uYf00\\_8pushSiYns29T\\_-ThI0qQaqpKbVS\\_LDqlAg/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1Q_uYf00_8pushSiYns29T_-ThI0qQaqpKbVS_LDqlAg/edit?usp=sharing)



- ▶ In FPGA, separate clusters into buckets
  - Data format in minimal chunk = one cluster in one channel: 2x10 bit header (channel ID + length) + 5x10 bit wavelet
  - Buffer long enough to allow transmission time spread, FVTX used 64 BCO buckets
  - Use internal memory on FPGA for BCO buckets storage (1.3kB \* 64 BCOs)
- ▶ **Max continuous rate/DAM = 115 Gbps, 2kB/BCO**
- ▶ Average continuous rate = 57 Gbps

# Throttling VS triggering

Rate estimation spread sheets:  
[https://docs.google.com/spreadsheets/d/1Q\\_uYf00\\_8pushSiYns29T\\_-ThI0qQaqpKbVS\\_LDqlAg/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1Q_uYf00_8pushSiYns29T_-ThI0qQaqpKbVS_LDqlAg/edit?usp=sharing)

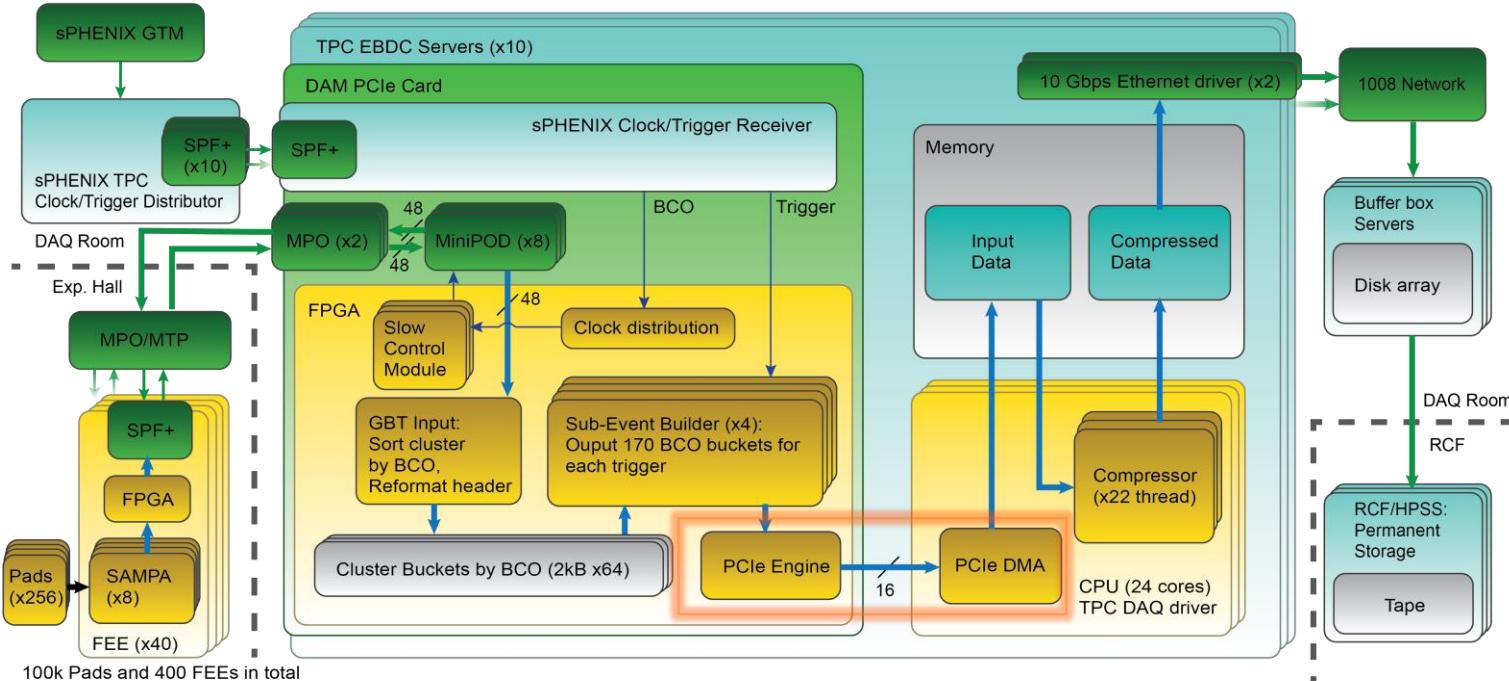


- ▶ 15kHz trigger + 170 BCO readout length (readout 18us data per trigger) → only need ~25% data from the input continuous stream
- ▶ Two options
  - Throttling: only record hits within 170BCO of the trigger and form a continuous data stream; no duplicated hits. **Data reduction to 25.5%**
  - Trigger: for each trigger, readout a chunk of hits timed to the next 170BCO. Form sub-event and easy for analysis; but could duplicate hits in output data if two trigger comes within 170 BCO. **Data reduction to 28.5%**
- ▶ Since the trigger mode only increase data volume by 10% (relatively), I would prefer trigger mode instead of throttled mode for easy analysis and monitoring.
- ▶ **Output max continuous rate/DAM = 33 Gbps**
- ▶ Output average continuous rate = 16 Gbps

# FPGA -> CPU

## Rate estimation spread sheets:

[https://docs.google.com/spreadsheets/d/1Q\\_uYf00\\_8pushSiYns29T\\_-ThI0qQaqpKbVS\\_LDqlAg/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1Q_uYf00_8pushSiYns29T_-ThI0qQaqpKbVS_LDqlAg/edit?usp=sharing)

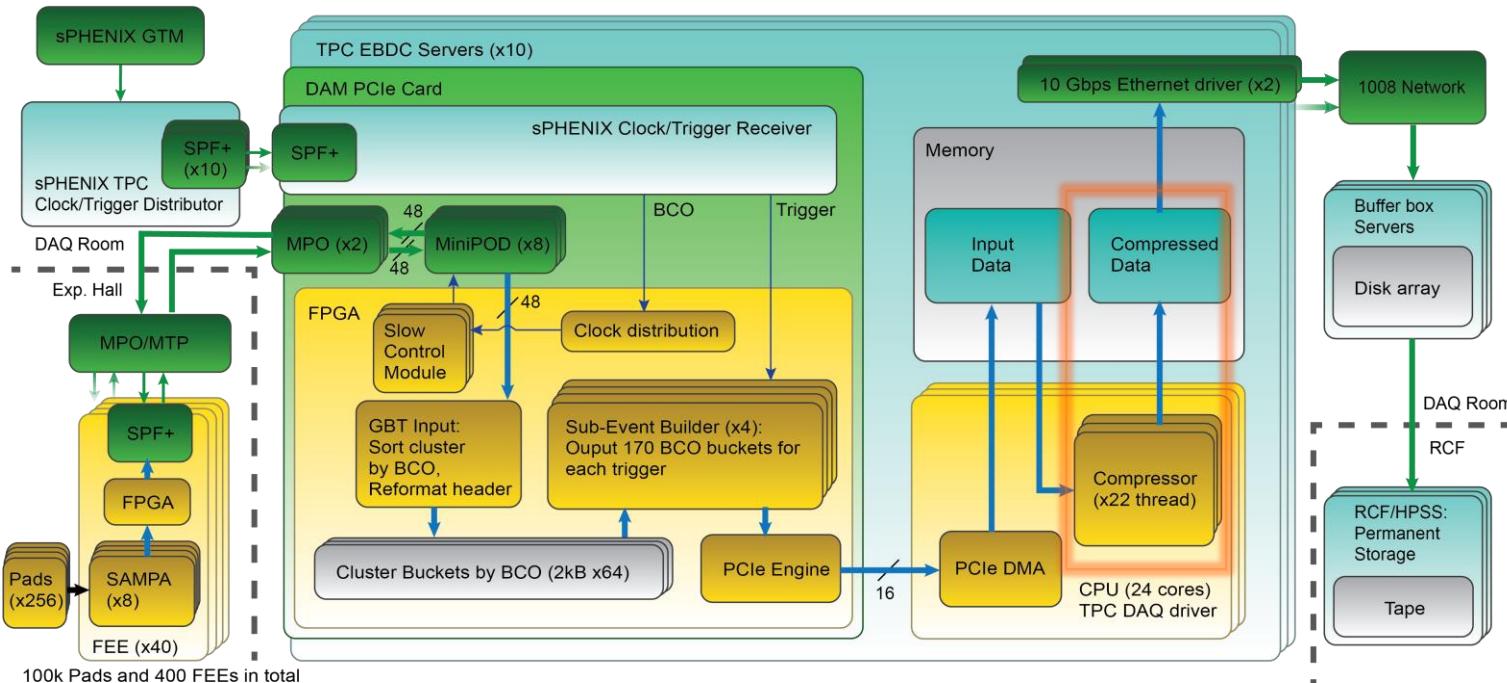


- ▶ FIFO and DMA event building output to Server Memory
  - Media: PCIe Gen3 x16
- ▶ **Demonstrated rate limit for FELIX (PCIe x16) = 100 Gbps**
- ▶ **Max continuous rate/DAM = 33 Gbps**
- ▶ Average continuous rate = 16 Gbps

# Data compression

## Rate estimation spread sheets:

[https://docs.google.com/spreadsheets/d/1Q\\_uYf00\\_8pushSiYns29T\\_-ThI0qQaqpKbVS\\_LDqlAg/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1Q_uYf00_8pushSiYns29T_-ThI0qQaqpKbVS_LDqlAg/edit?usp=sharing)



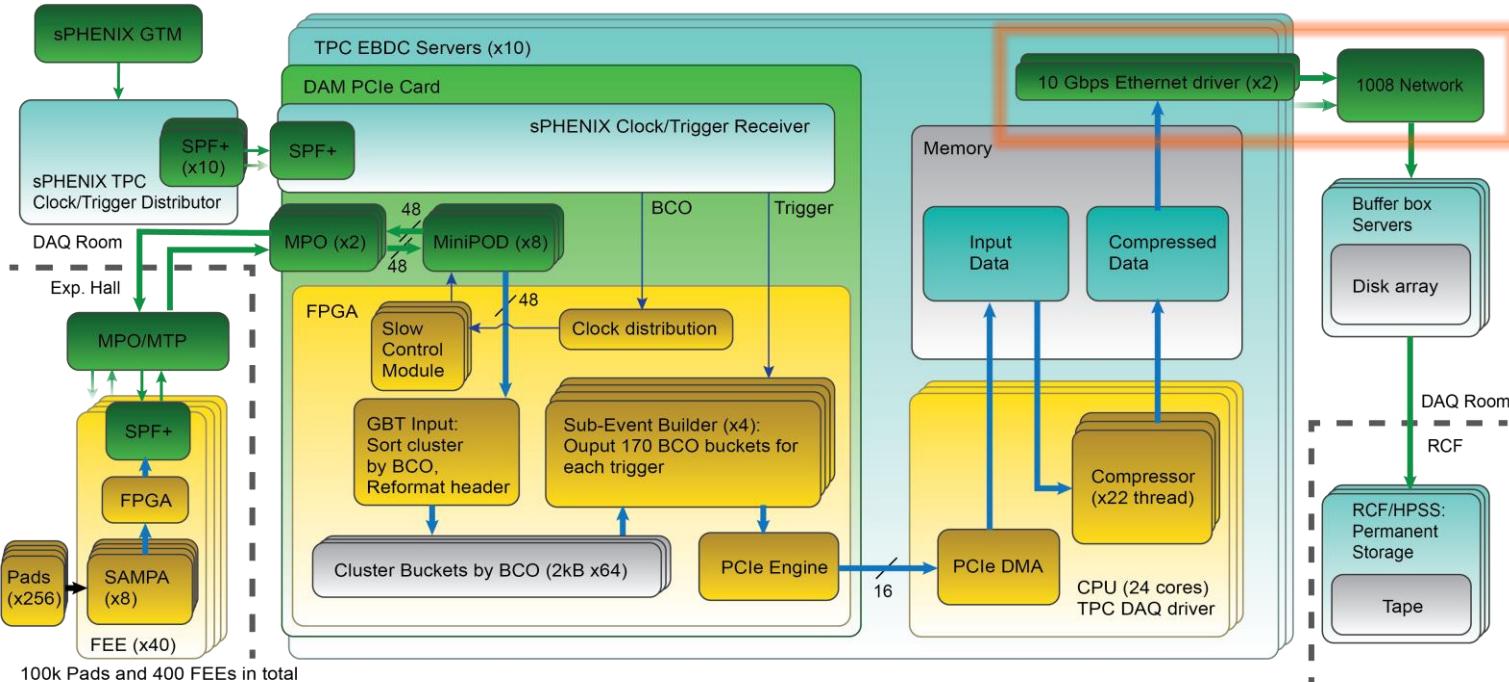
- ▶ Multithread compression
  - Algorithm: LZO on multi-event chunks
  - Demonstrated compression ratio = 60%
- ▶ **Estimated rate limit = 120 MBps / core = 21 Gbps**
- ▶ **Max continuous rate/DAM = 19.6 Gbps**
- ▶ Average continuous rate = 9.8 Gbps
- ▶ Backup option: Xilinx-based commercial FPGA code block run on gzip, 16k LUT, 100 Gbps  
<https://www.xilinx.com/products/intellectual-property/1-7aisy9.html#metrics>



# Output stage

## Rate estimation spread sheets:

[https://docs.google.com/spreadsheets/d/1Q\\_uYf00\\_8pushSiYns29T\\_-ThI0qQaqpKbVS\\_LDqlAg/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1Q_uYf00_8pushSiYns29T_-ThI0qQaqpKbVS_LDqlAg/edit?usp=sharing)

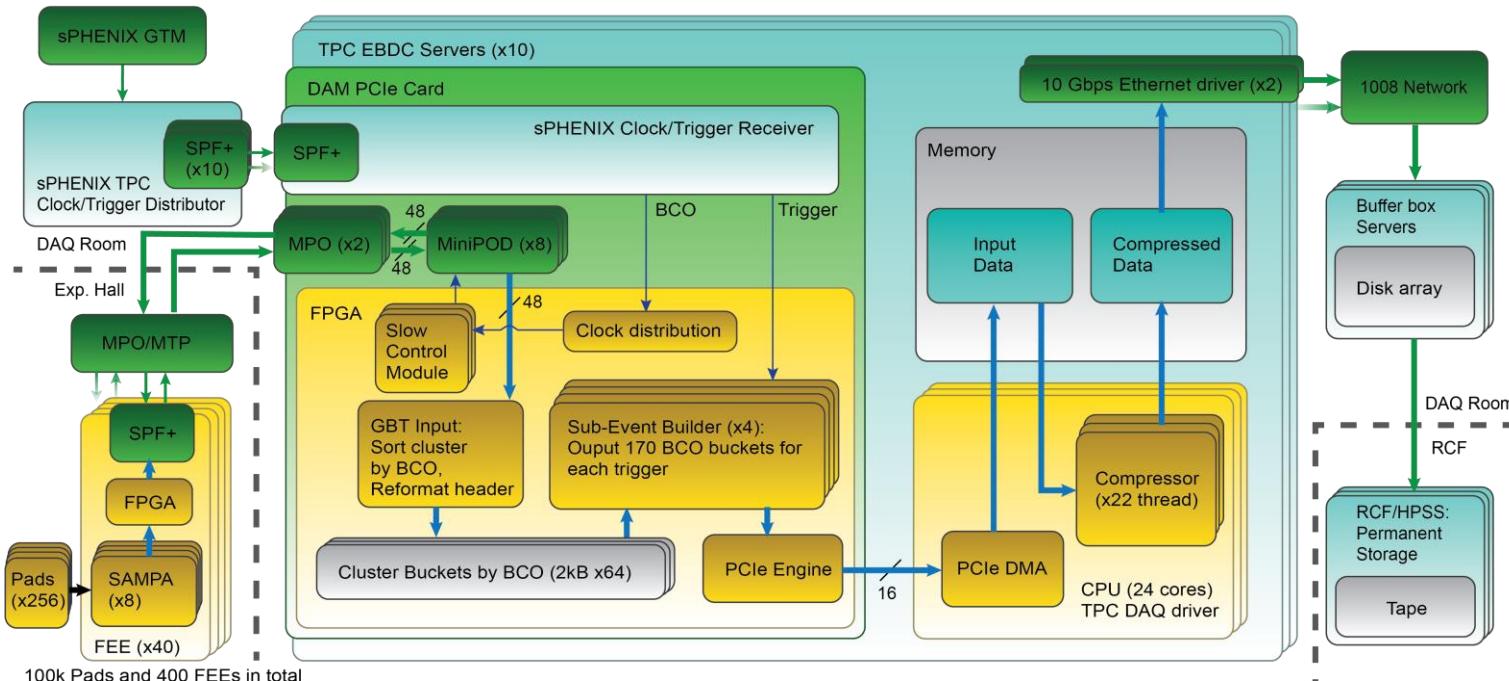


- ▶ Output to event builder
  - Media: 1x or 2x 10 Gbps Ethernet ports per EBDC server
- ▶ Rate limit media / EBDC server = **20 Gbps payload ?**
- ▶ Rate limit buffer box = **120 Gbps total? (3x HPSS rate)**
- ▶ Max continuous rate/EBDC = **19.6 Gbps**
- ▶ Average continuous rate for whole system = **98 Gbps**

# Summary

## Rate estimation spread sheets:

[https://docs.google.com/spreadsheets/d/1Q\\_uYf00\\_8pushSiYns29T\\_-ThlOqQaqpKbVS\\_LDqlAg/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1Q_uYf00_8pushSiYns29T_-ThlOqQaqpKbVS_LDqlAg/edit?usp=sharing)

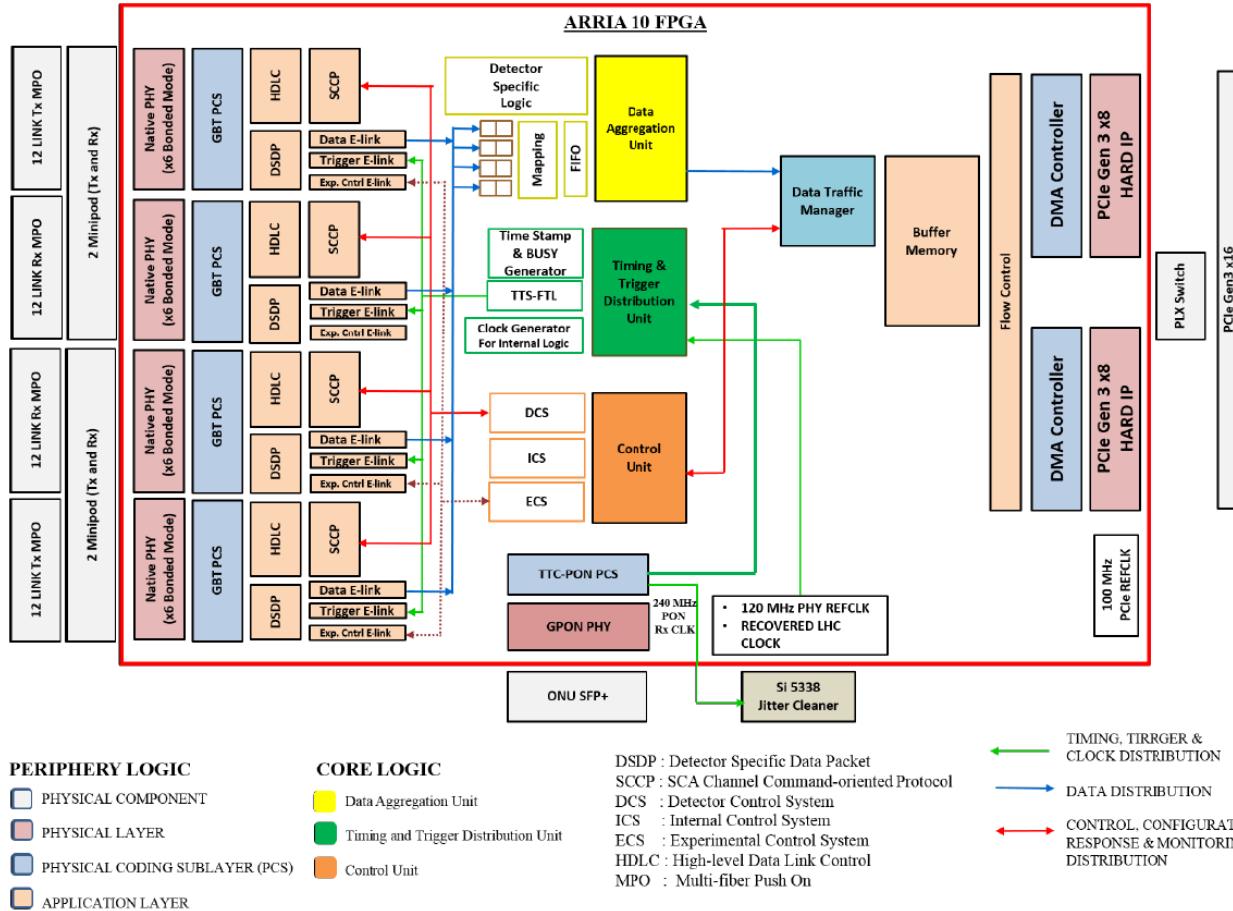


Item	Unit	Count	Sum Over all units				Per unit			
			Limit	Average	Max Continous	Max Instantious	Limit/Unit	Average/Unit	Max C./Unit	Max I./Unit
FEE SAMPA data	Gbps	3,200	4,096.00	573.44	1,146.88	3,870.72	1.28	0.18	0.36	1.21
FEE GBT fiber	Gbps	400	1,280.00	573.44	1,146.88	3,870.72	3.20	1.43	2.87	9.68
<hr/>										
FPGA Input	Gbps	10	1,280.00	573.44	1,146.88	3,870.72	128.00	57.34	114.69	387.07
Build hit - time table	Gbps	10	2,000.00	573.44	1,146.88	3,870.72	200.00	57.34	114.69	387.07
After triggering	Gbps	10	2,000.00	163.43	326.86	1,103.16	200.00	16.34	32.69	110.32
FPGA -> PCIe16 -> DMA	Gbps	10	1,000.00	163.43	326.86	1,103.16	100.00	16.34	32.69	110.32
Lossless Compression	Gbps	10	211.20	98.06	196.12	661.89	21.12	9.81	19.61	66.19
Server output to 1008 network	Gbps	10	200.00	98.06	196.12	661.89	20.00	9.81	19.61	66.19
<hr/>										
Disk server	Gbps	7	120.00	98.06				14.01		

# Extra Information



# CRU diagram



# SAMPA/STAR iFEE

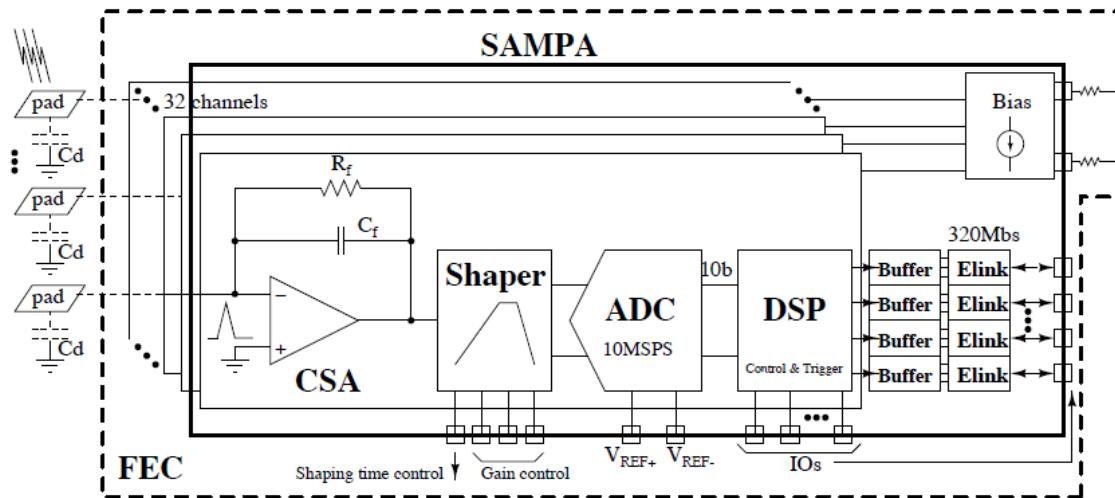
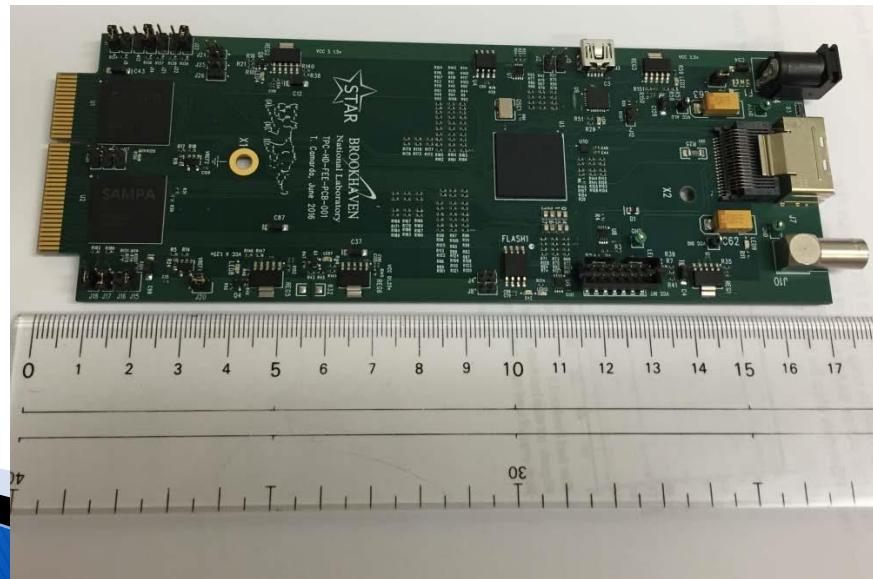


Figure 6.4: Schematic of the SAMPA ASIC for the GEM TPC readout, showing the main building blocks.



# Sept 2016 cost estimate

## Cost estimate (for production)

- Direct M&S cost for 100K channels is 1.1M FY16\$
- Cost for development is not included.
  - We assumed ~20-30% of this as M&S cost for development

Item	# of items	\$ per item	\$ all
SAMPA Chips	3200	\$44	\$140K
FEE cards	400	\$700	\$280K
DAM	50	\$6000	\$300K
Cables/fibers			\$100K
Power Supply	8	\$12000	\$100K
EBDC	50	\$3000	\$150K
Total			<b>\$1.1M</b>

c.f. STAR iFEE is \$150/card (64 ch., copper cable readout)